A Study on the Influencing Factors of International Used Sailboat Prices Based on SPSS Analysis

Maolin Yang, Xiaocui Du and Tong Li

School of Mathematics & Physics, Qingdao University of Science and Technology, Qingdao, China

Abstract. This paper focuses on the factors influencing the price of used sailboats internationally. First of all, this paper collected the sales prices of various types of used sailboats and obtained the indicators of each factor that may affect their prices. Due to the existence of multiple influencing factors, this paper decided to establish a multiple regression model to analyze the influence of each factor on the price. After conducting the r^2 test, it was determined that there was no linear correlation. Then a polynomial regression model was constructed by introducing a cross term. And it was continued using partial F test. We analyzed the data step by step and concluded that there is no linear relationship between the prices of the two sailboats and their effects, they correspond to a polynomial regression relationship and passed the model test. Afterwards, in order to use the mathematical model that has been developed to focus on the influence of region on the price of sailboats, this paper uses hierarchical clustering to initially classify the data, we use the average market price of each type of sailing in a given region as a measure of influence, and we compute the specific equation relationship based on the algebra of the model by comparing the coefficients of the equations to analyze the influence of regional factors on the price of the boat, and we obtain the The conclusion that the influence is statistically significant. Finally, we analyzed all the data and drew conclusions of specific practical significance.

Keywords: Multiple Regression Model; Polynomial Regression Model; Chi-square test; F-test.

1. Introduction

As with many luxury items, the value of sailboats varies with their age and market conditions. In order to better understand the sailboat market, we have explored the impact of various factors on sailboat prices through SPSS analysis and discussed their practical implications.

Firstly this paper develops a mathematical model to explain what factors influence the listing price of each sailboat in the market. At the same time, the resulting model is used in conjunction with a hierarchical clustering approach to focus on the influence of region on sailboat prices, and finally the practical implications of the model are discussed.

2. Data processing

Prior to modeling, this paper begins with a series of preparatory tasks, mainly data collection and data preprocessing. First we collected data on approximately 3,500 sailboats of 36 to 56 feet in length that were advertised for sale in December 2020 in Europe, the Caribbean, and the United States.

After initial sorting of the data, we found that there were many outliers. We chose to call the panda's library in python to fill in the data gaps, and then used Lagrangian interpolation to analyze the neighborhoods of the missing values, trimming the gap-filling values and removing the data that differed too much from the rest.

The new data obtained after the above process can then be used as valid data for problem analysis. We reviewed the data from many used sailboat sales websites and finally chose to use crawling techniques in Python software to collect data from the websites on the factors that affect the price of sailboats. We then chose four new metrics as additional data, including each country's breadth, draft, engine power, and GDP. Where each country's GDP was added to the known data, referred to as economic data. We used GDP data for each country from 2005 to 2019, covering the economic situation and time interpretation. Table 1 below shows a partial representation of the processed data.

Table 1. Presentation of results after data processing (partial).

Make	Variant	Length(ft)	Engine power(hp)	Beam(m)	Draft(cm)	Geographic Region	Country/Region/State	Listing Price(USC)	Year	GDP
Lagoon	380	38	20	11.55	115	Caribbean	Martinique	$204921	2005	0.067
Lagoon	380	38	20	11.55	115	Caribbean	Guadeloupe	$200071	2005	0.067
Lagoon	380	38	20	11.55	115	USA	Florida	$219000	2005	0.076
Leopard	40	39	29	6.72	135	Caribbean	Panama	$200000	2005	0.248
Broadblue	385	38.7	29	6.27	110	Europe	United Kingdom	$219923	2006	0.0509
Broadblue	385	38.7	29	6.27	110	Europe	Spain	$219766	2006	0.04

Secondly in order to explore the influence of regions on sailboat prices, we continue to use Python crawling techniques to collect and summarize coordinate data for each region. Table 2 below is a partial representation of the collected coordinate data.

Table 2. Coordinate data collected for each area.

	Coordinate data	
Country/Region/State	Longitude	Latitude
Alabama	-82.699	32.2668
Antigua and Barbuda	-61.8333	17.6167
Aruba	-69.9683	12.5211
Bahamas	-77.3963	25.0343
Belgium	4.4699	50.5039
Belize	-88.4976	17.1899
British Virgin Islands	-64.6399	18.4207
Bulgaria	25.4858	42.7337

3. An explanatory model of sailboat prices

3.1. Train of thought

In the resulting dataset, since GDP is able to override the year as a factor, we stopped considering the year and simplified the problem to consider only nine influencing factors. We need to develop a mathematical model that explains the price of each sailboat in the data table using the known influencing factors. Based on the fact that there are multiple influencing factors, we first consider a multiple regression model. Since we do not know how the variables are related to each other, we can construct a regression linear model and use matrix and least squares estimation to compute an estimate of β Methods [1]. Once this is done, we can perform an r^2 test on them to determine if they are linearly related to each other. If there is a linear correlation, the test is continued using the F-test, otherwise the operation is carried out using the partial F-test. In this process, by analyzing and processing the data step by step, we can get the conclusion we seek. For the estimation of the precision of the conclusion we choose to continue using the F-test method and use the F-value to determine whether the conclusion is precise and reliable.

3.2. The modeling process

We first develop a linear regression model of the following form.

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + ... + \beta_9 x_{19} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + ... + \beta_9 x_{29} + \varepsilon_2 \\ \vdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + ... + \beta_9 x_{n9} + \varepsilon_n \end{cases} \tag{1}$$

This linear regression model can be expressed in the form of a matrix as $\vec{y} = X\vec{\beta} + \vec{\varepsilon}$ .In this expression, there is the relation.

$$\vec{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},\ X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{19} \\ 1 & x_{21} & \cdots & x_{29} \\ 1 & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{n9} \end{bmatrix},\ \vec{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_9 \end{pmatrix},\ \vec{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{2}$$

We continue to use least squares estimation to solve for $\beta_0 \ldots \beta_9$ .At this point, the condition of minimum sum of squares of deviations needs to be met. The expression for the off-difference sum of squares is equation (3).

$$Q(\beta_0, \beta_1 \ldots \beta_9) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \ldots \beta_9 x_{i9} \right)^2 \tag{3}$$

So it is possible to derive the relation as follows.

$$Q(\beta_0, \beta_1 \ldots \beta_9) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \ldots \beta_9 x_{i9} \right)^2 = \min_{\beta_0 \ldots \beta_9} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \cdots \beta_9 x_{i9} \right) \tag{4}$$

The derivative of equation (4) yields equation (5).

$$\begin{cases} \left. \frac{\partial Q}{\partial \beta_0} \right|_{\beta_0 = \beta_0} = -2 \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \ldots \beta_9 x_{i9} \right) = 0 \\ \left. \frac{\partial Q}{\partial \beta_1} \right|_{\beta_1 = \beta_1} = -2 \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \ldots \beta_9 x_{i9} \right) = 0 \\ \vdots \\ \left. \frac{\partial Q}{\partial \beta_9} \right|_{\beta_7 = \beta_7} = -2 \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \ldots \beta_9 x_{i9} \right) = 0 \end{cases} \tag{5}$$

The resulting equation can be organized to obtain $X'(y - x\beta) = 0$, and then be able to get $X'(y - x\beta) = 0$, $\beta = \left(X'X \right)^{-1} X'y$ That is, there is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_9 x_9$. After completing this process, we have to determine whether there is a linear correlation between $x_1, ..., x_9$ and y_i . Here we use the r^2 -test method for judgment, where r^2 is calculated by the following formula.

$$r^2_{y_{1;2...,9}} = \frac{SSE(x_2, ..., x_9) - SSE(x_1, ..., x_9)}{SSE(x_2, ..., x_9)} \tag{6}$$

Algebraically, if r^2 = 0, it is shown that the linear relation does not exist and the larger the difference between its value and 0, the larger the correlation, and vice versa. If we obtain the result that there is a linear relation between the two variables, the partial F-test is further applied. Using the

stepwise regression method, the variables were introduced sequentially. For each independent variable introduced, the selected variables were tested one by one, and when the originally introduced variables were no longer significant due to the introduction of subsequent variables, they were excluded. F-tests were performed at each step. The formula for this step is as follows.

$$\begin{cases} \Delta SSR_{(j)} = SSR - SSR_{(j)} \\ F_j = \dfrac{\Delta SSR_{(j)} / 1}{SSE / (n - p - 1)} \end{cases} \tag{7}$$

If the calculation yields the result that the two do not have a linear relation, a n th-order polynomial regression model is further fitted in the following form.

$$\begin{aligned} y_i = {} & \beta_0 + \beta_1 x_{i1} + ... + \beta_9 x_{i9} + \beta_{11} x_{i1}^2 + ... + \beta_{i9} x_{i9}^2 + \\ & ... + \beta_{12} x_{i1} x_{i2} + ... + \beta_{19} x_{i1} x_{i9} + ... + \beta_{29} x_{i2} x_{i9} + ... + \beta_{49} x_{i4} x_{i9} \end{aligned} \tag{8}$$

Based on the above model, we can calculate the beta value. After calculating the partial F-value and eliminating the values that do not meet the F-test value, we can obtain the trend of correlation between each variable, and finally we can draw the requested conclusion. In the process of testing the accuracy of the results, we continue to use the F-value test to test the accuracy of the conclusions. In general, the smaller the F-value, the more accurate the results are and the higher the accuracy of the conclusions.

3.3. Model Solution

The model is solved using SPSS software [2]. First, the organized data were imported into the software and the r^2-values were calculated separately for the monohull sailboats prices as well as for the catamarans prices and their influencing factors. By reading and comparing the calculated r^2 -values, it is easy to conclude that there is no linear correlation between the two types of sailboat prices and their influencing factors. Table 3 below is a summary of the calculated R and the relations between the variables.

Table 3. Summary of r^2 and the correlation between each variable.

Monohull Sailboats			Catamarans		
Influencing Factors	Relations	r^2	Influencing Factors	Relations	r^2
x1	Secondary Relations	0.03	x1	S	0.003
x2	-	-	x2	TR	0.04
x3	logistic	0.317	x3	TR	0.494
x4	logistic	0.082	x4	S	0.231
x5	TR	0.33	x5	TR	0.003
x6	TR	0.032	x6	Secondary Relations	0.058
x7	Secondary Relations	0.005	x7	Secondary Relations	0.004
x8	TR	0.007	x8	TR	0.003
x9	power	0.017	x9	TR	0.221

Intuitively, from the table we can conclude that there is no linear relationship between the variables and therefore an nth order linear regression model was fitted. The values of β and F were calculated according to the solution formula. We obtained results consistent with the above findings, thus further calculating and fitting the F-values. The above process is shown in Fig. 1, with data from a monohull sailboat on the left and a catamaran on the right. The images in the table are visualizations of the calculation and fitting process. We can see that the F-value is relatively small and fixed in the interval (0, 1). This shows that the conclusions drawn are more accurate and reliable.

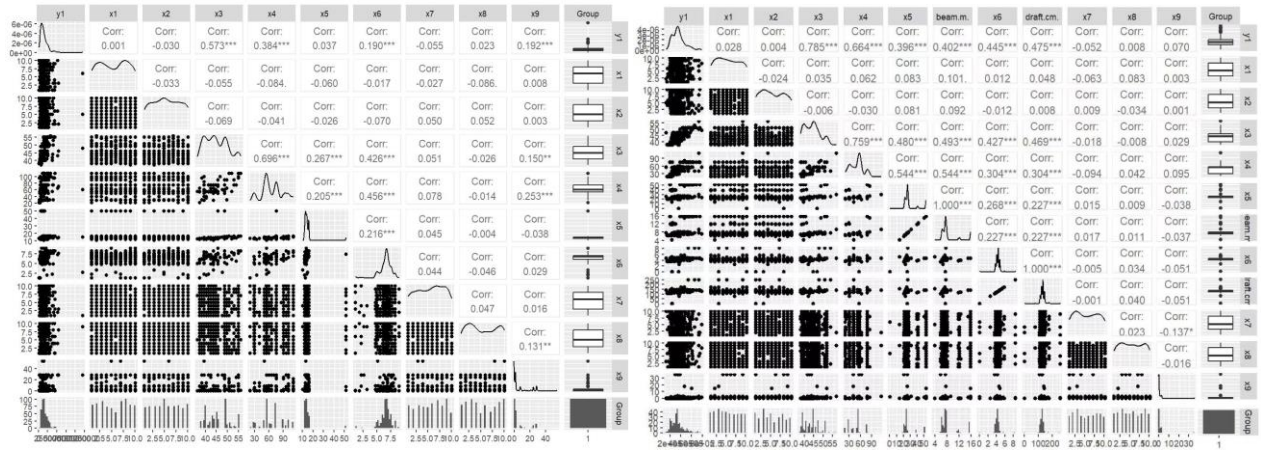

Fig 1. Visualization of the data fitting process.

4. A study of regional influences on sailboat prices

4.1. Train of thought

In order to explore the influence of regions on sailboat prices, we chose to use a hierarchical clustering approach [3], the idea of which is shown in Fig. 2.

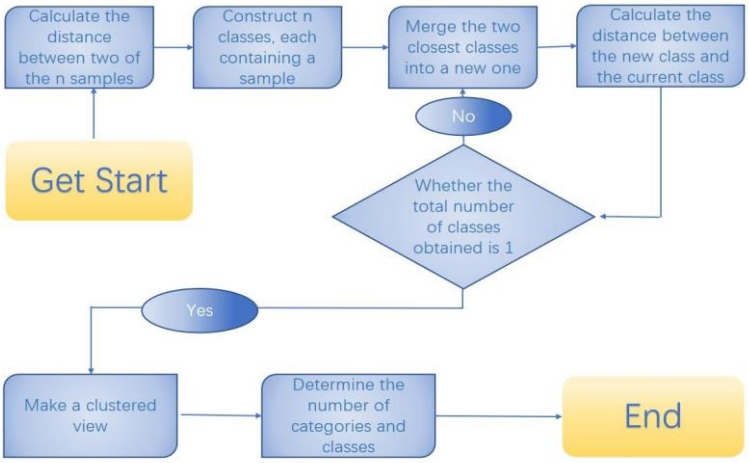

Fig 2. Mind map of the hierarchical clustering approach.

The specific method of clustering consists of five main steps. First, each area object given in the data is considered as a subclass, then the distance between each area is calculated using the collected coordinate data and compared in turn, the area with the smallest distance from other objects is put into a class, and then the above operation is repeated for the new class objects until all the class objects are put into one big class.

We need to find the average market price of each sailboat model in a given region and use it as a measure of the impact of regional factors on the price of sailboat models. Based on this model, specific equations can be derived algebraically to be able to analyze the impact of regional factors on sailing variants. Determining consistency requires a consistency test for the consistency coefficients. If the consistency coefficients pass the consistency test, the regional factors for the variable can be considered consistent. The next thing that needs to be analyzed is the statistical significance and we use the F-value address to derive the result by comparing it with the critical value. Finally the practical significance is summarized by analyzing all the data.

4.2. Modeling process

After hierarchical clustering, we assume that all the regional objects are finally classified into k classes. We are able to obtain a regression model in the form of the following equation.

$$\overline{y}_i = \beta_0 + \beta_1 a_i, i = 1,...,k \tag{9}$$

$\overline{y}$ Is the average market price of used sailboat variants in each region? The market price is found by first fixing each sailboat variant and then fixing the regional factor. Then proceed to calculate the sum of the data. We can obtain the total market price data for each regional sailboat variant. At the end of the above steps, we get the total data. We proceed to sort the categories made up of regional data, roughly by the total price at which the sailboat variant sells in each regional category. The sorting rules are as follows: the category with the lowest total sale price is numbered 1, in ascending order, from smallest to largest. We visualized the collected data and finally obtained Fig. 3 below, which shows the approximate impact of region on the market price of sailboat variants.

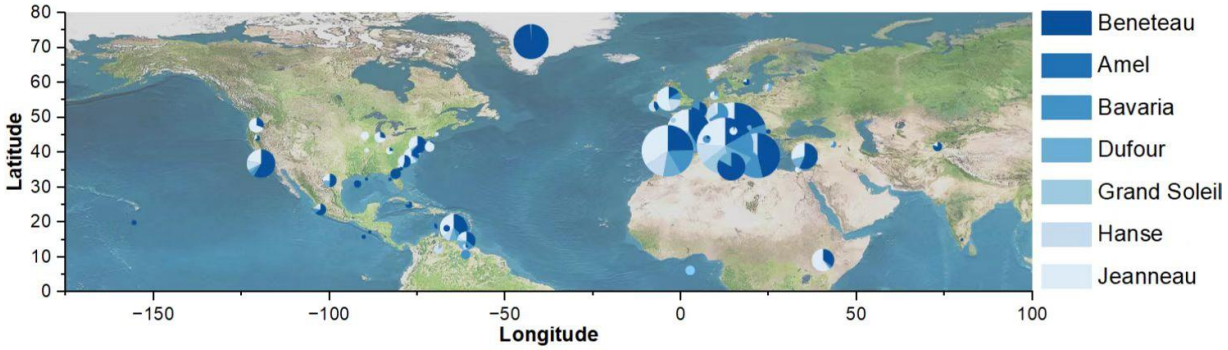

Fig 3. Collected data after visualization.

The stepwise process of fixing the variables to find the data utilizes SPSS, and the following Table 4 presents a partial presentation of the total data. [4]

Table 4. Partial presentation of the total data.

Total Data(Take calamarns as an example)				
Number	Average	Make	Variant	Country/Region/State
43	534,500	Fountaine Pajot	Helia 44	Bulgaria
10	484,286	Fountaine Pajot	Helia 44	Florida
50	445,750	Fountaine Pajot	Helia 44	France
48	552,319	Fountaine Pajot	Helia 44	Greece
51	453,111	Fountaine Pajot	Helia 44	Grenada
41	559,900	Fountaine Pajot	Helia 44	Italy
49	518,171	Fountaine Pajot	Helia 44	Martinique
46	430,705	Fountaine Pajot	Helia 44	Maryland
35	589,000	Fountaine Pajot	Helia 44	Spain
45	479,801	Fountaine Pajot	Helia 44	U.S. Virgin Islands
43	582,000	Fountaine Pajot	Helia 44	Netherlands

Our solution is to target each sailboat variant in turn and find the relation between each type of region and the market price of that sailboat variant. The data is then analyzed to introduce the type of relation that is presented between the market price of the sailboat variant and the regional factors. In turn, we are able to further analyze the data and fit it to the known relations to draw conclusions.

For the consistency determination, we need to perform consistency tests for the effects of the regional factors analyzed above for each sailing variant. First, the consistency coefficient W needs to be calculated. The steps are as follows:

Step 1: Find the average value $\overline{y_i}$ of the prices of all sailboat variants.

Step 2: Calculates the sum of the squared differences between the listing price and the price average. Here is the formula:$s = \sum_{i=1}^{n}(\hat{y} - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - n\overline{y}$

Step 3: The consistency coefficient is calculated by substituting the derived data into the formula. Here is the formula: $W = \frac{12s}{k(n^3-n)}$

After obtaining the consistency coefficients, we tested the consistency coefficients using the χ^2 test.χ^2 Is calculated as$\chi_0^2 = k(n-1)W$. We choose the case where the confidence level α is equal to 0.05 .Our method of determination is as follows: If $\chi_0^2 > \chi_\alpha^2(k-1)$ holds, the consistency coefficient is considered to successfully satisfy the test.

For the practical and statistical significance of any regional, we chose to address this using the F-value. The formula to calculate the F-value is below.

$$F = \frac{s^2_1}{s^2_2} = \frac{\frac{\sum\left(X-\bar{x}\right)^2}{n-1}}{\frac{\sum\left(X-\bar{x}\right)^2}{n-1}} \tag{10}$$

Here we choose the case where the significance level is 0.05, and by consulting the F distribution table and comparing the actual derived F values with the found critical values, we can finally draw a conclusion.

4.3. Model Solution

By analyzing the data, we determined that the market price of the sailboat variant showed a polynomial regression relation with regional factors. We used SPSS to solve the equation to derive the coefficients B_1 and B_2. The specific equation data are shown in Table 5 below.

Table 5. Specific data of the equations.

	$y = Intercept + B_1x_1 + B_2x_2$	
	B_1	B_2
Number	16914.71407±1068.69619	209.70761±24.30609
	16044.9395±2992.14313	-199.34388±65.35382
	34348.5034±3752.6424	505.0080±84.03301
	.....	
	4778.54743±823.36689	-50.89381±15.75218
	14079.99295±1764.69619	-149.0541±28.41454
	13142.54985±1627.90905	-151.47256±26.0221

Analyzing the specific data of the equation, we can conclude that the trend of the regional effect on the listing price of sailboats can be simply described by the coefficients of the equation and, based on the specific values of the coefficients, we are able to make quantitative inferential calculations of the listing price using regional factors.

After further analysis of the data, we chose to continue the fitting operation according to the polynomial regression relation, and the fitting procedure is shown in Fig. 4 below.

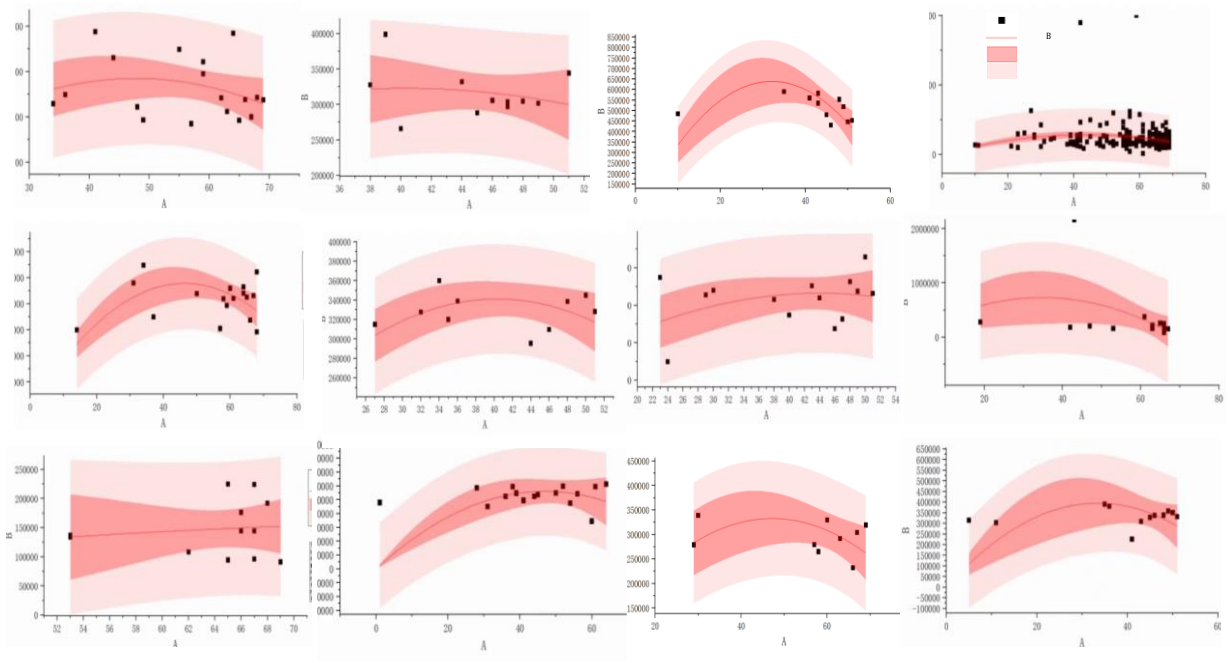

Fig 4. Schematic diagram of the fitting process.

We calculate a series of data required for the chi-square test and combine the data summary table for monohull sailboats and catamarans as shown in the following Table 6. [5]

Table 6. Specific data of the equations.

Consistency Check Data Sheet		
Variables	Numerical value	
$\bar{y}$	Monohull sailboats	catamarans
	255329	472468
s	1.3×10^{13}	2.6242×10^{13}
w	9966.6905	1972.60016
χ_0^2	909141574	65.1708

According to the figures, when the confidence level is 0.05, the comparison of the Chi-square test for monohulls and catamarans is as follows.

$$\chi^2_{0.05}(69-1)=88.12502>\chi^2_{0.05}(69-1)$$
$$\chi^2_{0.05}(49-1)=65.17080>\chi^2_{0.05}(49-1) \tag{11}$$

It is clear that the consistency coefficient of the influence of regional factors on the market price of used sailboats passes the consistency test, and we can consider it to be consistent.

For significance, we can calculate the F-value by using the formula of F. Some of the data of F-value are presented in Table 7 below. At the set significance level α equal to 0.05, we consult the F-distribution table to obtain$F(-\infty,+\infty)$, which allows us to obtain the critical value of the F-value. Comparing the actual derived F-values with the critical values, it is easy to find that each of the f-values is much larger than the critical value. Therefore, it is concluded that the regional effect is significantly different for all variants. We are able to infer that the regional effect is statistically significant for the listing price of sailboats.

Table 7. Partial summary table of the F –value.

F-value section summary table						
x_{i1}	x_{i2}	x_{i3}	.....	x_{ik-2}	x_{ik-1}	x_{ik}
y_i	171.15044	6.21379	248.42203	.....	167.76891	655.41083 182.6783

Looking back at the process of analysis we completed above and the data obtained, we can summarize the practical significance of the regional effect on the listing price of used sailboats. In the context of global economic development, it is important to study the difference in commodity prices around the world. This is not only a test of the economic development capacity of a certain region, but also provides a credible sample for studying the characteristics of economic development. In particular, using real data from each region as an example can deepen everyone's understanding of economic development and cultivate an individual's vision of data forecasting.

5. Summary

In this paper, we have collected data on the listing prices of used sailboats and analyzed them in depth. We obtained a model of the relationship between used sailboats and various factors that can be used for price prediction of used sailboats, and the price prediction model can be used not only for determining the price of a sailboat for sale, but also for evaluating the reasonableness of the purchase price when buying a used sailboat. At the same time, through the qualitative analysis of certain indicators, we have also identified the region as an important factor affecting the price of sailboats, and this method can also be extended to other influencing factors to determine the main factors affecting the price of second-hand sailboats, to provide a theoretical basis for the buyers and sellers of second-hand sailboats.

References

[1] Yuhui Cai. Analysis of the Influencing Factors of China's Foreign Exchange Reserves Based on Multiple Linear Regression Models [J]. Trade Fair Economy, 2022(16):61-63.DOI:10.19995/j.cnki.CN10-1617/F7.2022.16.061.

[2] Junshuang Huo, Cong Chen, Qingsong Wang. A study on pricing of refined oil products based on SPSS software and regression analysis [J]. Electronics World, 2015(24):49+53.

[3] Zhihao Chen, Jingmin Ji. Compositional analysis and identification of ancient glassware based on hierarchical clustering models [J]. Modern Information Technology, 2023, 7(08):122-125.DOI:10.19850/j.cnki.2096-4706.2023.08.031.

[4] Chunyan Pan. Implementation of stepwise regression in statistical software and empirical analysis [J]. Science & Technology Vision, 2019(18):41-42.DOI:10.19694/j.cnki.issn2095-2457.2019.18.020.

[5] Chen Y,Nie J,Xu K. Classification and Prediction of Ancient Glass Artifacts based on Chi-square Test[C]//Wuhan Zhicheng Times Cultural Development Co., Ltd.. Proceedings of 2023 International Conference on Mathematical Modeling, Algorithm and Computer Simulation (MMACS 2023).Proceedings of 2023 International Conference on Mathematical Modeling,2023:272-277.DOI:10.26914/c.cnkihy.2023.008051.